How Does Alignment Tuning Shape Representations of Sycophancy and Related Cue-Induced Biases in LLMs?

Anonymous ACL submission

Abstract

Modern LLMs are alarmingly susceptible to surprisingly simple immaterial changes of input prompts: a casual hint, an incorrectly labeled few-shot example, or a fake prior assistant turn often flips an originally correct answer. We study where this susceptibility, spanning sycophancy and related cue-induced biases, lives inside the model. Across five model families and seven BCT bias types, we extract a per-bias direction from hidden states and triangulate it through three measures: probing, leave-one-dataset-out transfer, and causal intervention. The susceptibility is largely installed by alignment tuning rather than pretraining: pretrained base models barely cave to these biases, and their activations carry no cue-specific signal beyond question content. Within aligned models, each bias becomes a single coherent direction that we can both decode and steer along, recovering the unbiased answer across every family we test. The biases stay representationally distinct, however: cross-bias entanglement is model-specific rather than a property of the bias category, and even behaviorally similar biases occupy different directions. The same intervention also serves as a modest debiasing tool, recovering 7–20% of bias-induced errors while preserving $\geq$90% of correct answers across all five instruct families. Cue-induced bias is therefore best understood not as a single flaw in LLMs but as a family of distinct, causally active directions that alignment tuning installs.

1 Introduction

Modern LLMs are remarkably brittle to how an input question is asked. Inserting an offhand "I think the answer is (C)" before the question or a fake prior assistant turn often flips a model's answer on questions it would otherwise get right (Turpin et al., 2023; Sharma et al., 2024; Perez et al., 2022). Chua et al. (2024) cataloged a benchmark spanning sycophancy, distractor content, few-shot patterns, and post-hoc commitment (the BCT benchmark) and demonstrated the susceptibility behaviorally. The consequence is practical: in any setting where prompts are not tightly controlled—agentic systems, multi-turn assistants, search interfaces, automated evaluations—an LLM's answer can be quietly steered by these *cue-induced biases*.

Behavioral characterization tells us that this happens. It does not tell us where the susceptibility lives inside the model. Recent interpretability work has shown that single behaviors (refusal, sycophancy, true/false belief) can be isolated as linear directions in hidden-state space and steered along causally (Panickssery et al., 2024; Marks and Tegmark, 2023; Zou et al., 2023; Li et al., 2023). But three questions remain open about cue-induced biases specifically: (i) whether the susceptibility is already present in the pretrained network or is installed by alignment tuning, (ii) whether the per-bias direction construction generalizes across families and across bias types, and (iii) whether different biases in the same model share internal structure or remain separate.

We answer all three by extracting a direction for each (bias, dataset) source as the normalized difference of means between hidden states where the model caved to the bias and those where it resisted, the standard contrastive-activation-addition construction (Panickssery et al., 2024; Marks and Tegmark, 2023). The direction simultaneously functions as a probe (project to predict flipping) and a steering vector (subtract to suppress the bias, add to amplify it). For the origin question, base models pose a problem because they almost never flip and so produce no within-model flipped/resisted contrast to learn from. We solve this with a base-probe protocol that borrows the aligned model's flipped/resisted labels and asks whether the same distinction is decodable in the base model's activation space (§6).

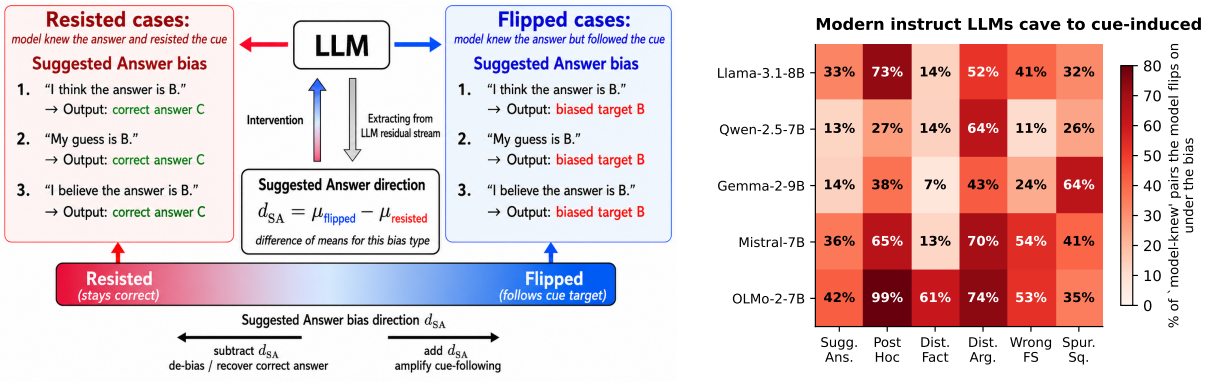

Figure 1: **Left: per-bias direction extraction and intervention.** For each (biased, unbiased) prompt pair, the model either *resists* the cue (gives the correct answer) or *flips* (caves to the bias's target letter). We extract $d_{\text{bias}} = \mu_{\text{flipped}} - \mu_{\text{resisted}}$ from last-token residual-stream activations (here illustrated for *Suggested Answer*); subtracting $\alpha \cdot d_{\text{bias}}$ at inference time debiases the model (§8), adding amplifies cue-following. **Right: scope of the problem.** Per-(model, bias) flip rate: fraction of pairs the model would have answered correctly on the unbiased prompt but flips on when reframed with the bias cue. OLMo's *Post Hoc* (99%) leaves no resisted contrast for LODO testing (§5).

We study **seven BCT bias types**[1] across **five model families** (*Llama-3.1-8B*, *Qwen-2.5-7B*, *Gemma-2-9B*, *Mistral-7B-v0.3*, *OLMo-2-7B*), using both the base and instruct checkpoints of each. The story that emerges has three parts.

Origin. Four of five pretrained base models flip on under 5% as many pairs as their instruct counterparts, and in their activations the cue-specific bias signal is essentially absent once we control for question content (§6). The biases that are not instruction-like (distractor content, few-shot patterns, post-hoc anchoring) are more absent from base activations than the one instruction-like bias (*Suggested Answer*), arguing against the simpler reading that we are just measuring sensitivity to instructions. Alignment tuning, not pretraining, is where the susceptibility comes from.

Per-bias direction. Across 7 biases and 5 instruct families, the within-bias direction transfers to held-out datasets at per-model mean AUROC of 0.69–0.82, with every test point beating a random direction at $p < 10^{-6}$ (§5). The signal also localizes consistently: when we extract the direction one layer at a time, AUROC peaks in the late-middle layers (relative depth 0.55–0.74) in every architecture we test, with the single best layer beating the all-layer probe.

[1] *Suggested Answer*, *Distractor Argument*, *Distractor Fact*, *Wrong Few-Shot*, *Spurious Few-Shot Squares*, *Spurious Few-Shot Hindsight* (singleton), and *Post Hoc*. We exclude *Positional Bias* (LLM-as-judge format) and *Are You Sure* (5-turn structure with an embedded reasoning paragraph incompatible with our non-CoT setup). Detailed definitions are in Appendix A.1.

Distinctness. The tight cross-bias "cluster" visible on Qwen is a Qwen-specific quirk, moderate on Gemma, weak on Mistral, and absent on Llama and OLMo (§7). And even behaviorally similar biases need not share a representation: *Post Hoc* and *Suggested Answer* are both answer-anchoring on the surface but occupy different directions inside the network, and are anti-aligned on Llama.

The same direction also doubles as a debiasing knob (§8): at a per-model strength chosen to keep correct answers correct ($\geq 90\%$), subtracting the direction from the residual stream recovers 7–20% of bias-induced errors across all five instruct families, versus under 5% for a random direction at matched strength.

2 Related Work and the Research Gap

Behavioral characterization is well-established, mechanistic characterization is not. Chua et al. (2024) introduced BCT and demonstrated that LLMs cave to a range of prompt manipulations spanning sycophancy, distractor content, few-shot patterns, and post-hoc commitment. Sycophancy in particular has been studied extensively (Sharma et al., 2024; Perez et al., 2022), and Turpin et al. (2023) showed chain-of-thought explanations are often unfaithful to the underlying decision. This line of work tells us that models exhibit the behavior. It does not address where the susceptibility lives inside the network or whether it differs across model families.

Previous interpretability work is largely on one behavior/model. A growing literature shows that single behaviors can be isolated as linear directions

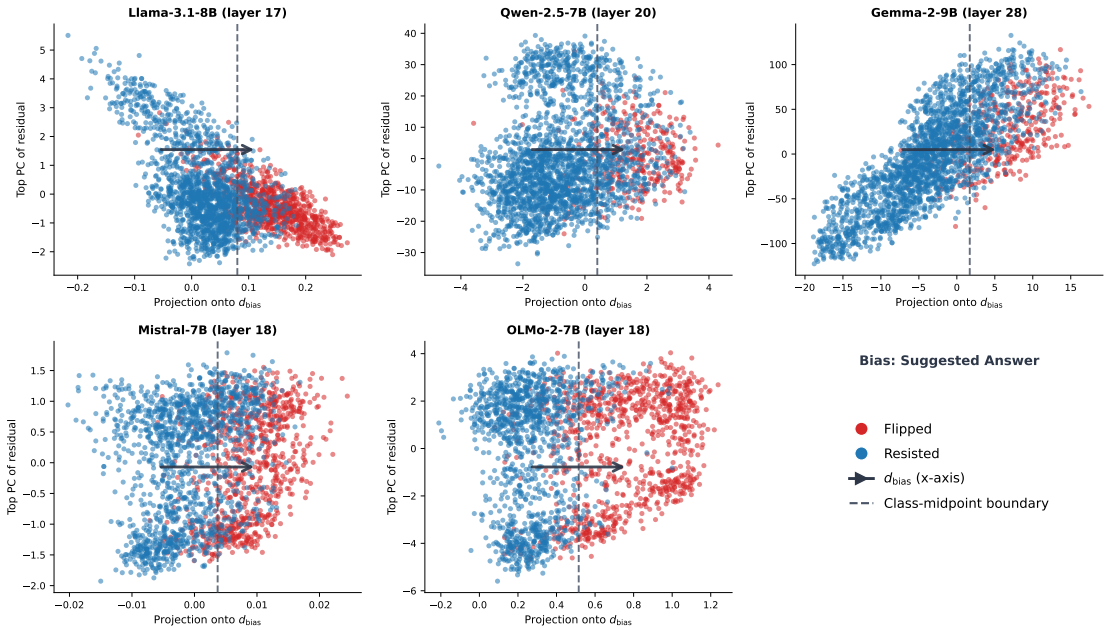

Figure 2: **The *Suggested Answer* direction separates flipped from resisted at the data level, in every family.** 2D view of last-token hidden states (4 datasets pooled per panel, at each model's *Suggested Answer*-specific best layer; Appendix B.1). X-axis: projection onto the per-bias d_{bias} (Fig. 1); y-axis: top PC of the residual orthogonal to d_{bias}. Flipped red, resisted blue; black arrow marks the d_{bias} axis, dashed line is the midpoint of class means along it. Held-out LODO AUROCs for *Suggested Answer* are in Table 1 (0.76–0.88).

in hidden-state space. Panickssery et al. (2024) extract a per-persona steering direction (Contrastive Activation Addition). Marks and Tegmark (2023) show that true/false statements lie along a linear direction. Hendel et al. (2023) find that in-context tasks are compressed into single task vectors. Li et al. (2023) show inference-time intervention can elicit truthful answers. Burns et al. (2022) discover latent knowledge unsupervisedly. Templeton et al. (2024) extract interpretable features via sparse autoencoders, and Zou et al. (2023) frame representation engineering more broadly. These works establish that per-behavior directions exist. They typically focus on one behavior in one model and do not ask whether the same construction generalizes across a benchmark of related biases, whether multiple biases share representational structure, or whether the resulting directions are inherent to the pretrained network or installed by alignment tuning.

The gap we fill. No prior work, to our knowledge, has systematically asked, across multiple bias types and multiple model families: (i) where the per-bias direction comes from (pretraining vs. post-training), (ii) whether different biases occupy the same or distinct directions, and (iii) whether the direction construction holds universally across model families. Filling this gap matters. If per-bias probing methods claim selectivity without verifying it cross-model, the field is overstating what those probes measure, and if alignment tuning is the source of the susceptibility, that has direct implications for safety-tuning recipes.

Our contribution. A systematic 5-family $\times$ 7-bias triangulation (probing + LODO transfer + causal intervention) plus a new base-probe protocol that compares base-vs-instruct *representations* (not just behaviors). The result is a three-part picture (origin, category, distinctness) that re-centers what "cue-induced bias" means inside the model.

3 Methodology

Data and bias types. We study the **7 BCT biases** (Chua et al., 2024) that fit a single-token MCQ pipeline: *Suggested Answer*, *Distractor Argument*, *Distractor Fact*, *Wrong Few-Shot*, *Spurious Few-Shot Squares*, *Spurious Few-Shot Hindsight* (a singleton with only one source dataset), and *Post Hoc*. We exclude *Positional Bias* (LLM-as-judge format, not MCQ) and *Are You Sure* (a 5-turn dialogue with a baked-in fake reasoning paragraph). Each non-singleton bias has four source datasets (MMLU, HellaSwag, LogiQA, TruthfulQA). Detailed bias definitions are in Appendix A.1.

Non-CoT, single-letter prompting. Throughout the paper we work in a *non-CoT* setup: instead of letting the model generate a chain-of-thought, we modify the user message that contains the CoT instruction (the final user turn in every BCT prompt) to ask for a direct single-letter answer (appending: "*Answer with just the single letter (A, B, C, D, ...), no explanation or reasoning.*"), forward-pass the prompt, and read the next-token logits over the 16 letter tokens A–P, summing probability mass across bare-letter and space-prefixed variants. This isolates the bias's effect on the answer itself, separate from any reasoning the model would produce.

Flipped vs. resisted pairs. For each (biased, unbiased) pair, let t be the bias target letter, c the ground-truth letter, a_b the model's argmax letter under the biased prompt, and a_u under the unbiased prompt. A pair is **flipped** if $a_u = c$ and $a_b = t$ (the model knew the answer but caved to the bias's letter) and **resisted** if $a_u = a_b = c$ (the model held the right answer). We drop pairs where $t = c$ (the bias coincidentally points to the truth) and pairs where the model did not know the unbiased answer. Pair-count statistics are in Appendix A.2.

Per-source and per-bias directions. For each (bias, dataset) source s, we collect the last-token hidden state of every transformer layer on each biased prompt, concatenated into $X_s \in \mathbb{R}^{LH}$. The per-source direction is the standard difference-of-means construction (Panickssery et al., 2024; Marks and Tegmark, 2023):

$$d_s = \frac{\mu_s^{\text{flip}} - \mu_s^{\text{rsst}}}{\|\mu_s^{\text{flip}} - \mu_s^{\text{rsst}}\|_2}. \quad (1)$$

The per-bias direction d_B is the unit-normalized mean of d_s across B's source datasets, treating each dataset equally.

Probing AUROC. Given a direction d and held-out pairs $\{(X_i, y_i)\}$ (flipped vs. resisted), we score $s_i = X_i \cdot d$ and report AUROC. The probing random baseline is the mean AUROC of $K = 50$ random unit-norm directions of matched dimension (per-layer probing uses $K = 20$, Appendix B.1). The intervention random baseline (§5, §8) uses a single random unit-norm direction per source (seed 42) at the same α as the real direction.

Base-probe protocol. The origin question (§6) cannot be asked directly on base models, since they rarely flip and so produce no within-model flipped/resisted contrast to learn a direction from. We instead *transfer* labels: for each source, we identify which pairs the *aligned* model flipped or resisted on, feed the base model the exact same prompts, record the base activations, and ask whether the aligned-model-defined distinction is decodable in base-activation space. To separate a real cue-specific signal from question-content variability (perhaps the aligned-model-flipped questions are simply harder), we run the analogous probe on the *unbiased* version of each prompt as a control and report the paired biased-minus-unbiased AUROC gap per source.

4 Experimental Setup

Models. Five instruction-tuned 7–9B models from different families: *Llama-3.1-8B-Instruct* (Grattafiori et al., 2024) (32L/4096H), *Qwen2.5-7B-Instruct* (Yang et al., 2024) (28L/3584H), *Gemma-2-9B-it* (42L/3584H, bf16 + eager attention required to preserve attention soft-capping), *Mistral-7B-Instruct-v0.3* (32L/4096H), and *OLMo-2-1124-7B-Instruct* (32L/4096H, bf16). For the origin experiment (§6) we also extract activations from each family's matching pretrained base checkpoint (*Llama-3.1-8B*, *Qwen2.5-7B*, *Gemma-2-9B*, *Mistral-7B-v0.3*, *OLMo-2-1124-7B*), each reusing its instruct tokenizer so that base-vs-instruct prompts are byte-identical and the only differing factor is the weights. *Qwen2.5-7B-base* is the exception that caves to biases at instruct-like rates and is therefore also kept in the main analyses as a behavioral contrast.

Datasets. MMLU (Hendrycks et al., 2020), HellaSwag (Zellers et al., 2019), LogiQA (Liu et al., 2020), TruthfulQA (Lin et al., 2022), plus Hindsight Neglect (McKenzie et al., 2023) for the singleton bias.

Decoding. Single-token greedy. We extract only the last-token hidden state at the answer position and the next-token logits over the 16 letter tokens A–P. No generation beyond a single token.

5 Each Bias Is a Single Causal Direction

Before we can ask where the bias comes from or how multiple biases relate, we first need to show the per-bias direction is a real, generalizable object. Across 7 biases and 5 instruct families, three measures agree. The direction decodes the bias from held-out data, it causally controls the answer when

Bias	Llama	Qwen	Gemma	Mistral	OLMo
Sugg. Answer	0.83	0.87	0.84	0.76	**0.88**
Post Hoc	**0.87**	0.86	0.83	0.71	—
Dist. Fact	0.75	0.81	0.78	0.72	0.75
Dist. Argument	0.77	0.80	0.74	0.64	0.80
Wrong Few-Shot	0.65	0.83	0.73	0.60	0.70
Spur. Squares	0.61	0.75	0.75	0.70	0.65
Overall mean	0.74	0.82	0.78	0.69	0.76
Random baseline	0.50	0.51	0.50	0.50	0.50

Table 1: **The direction generalizes across datasets in every family.** Mean LODO AUROC. Bold marks each model's strongest bias. "—" = OLMo's *Post Hoc* cannot be LODO-tested. All non-blank cells beat random at $p < 10^{-6}$.

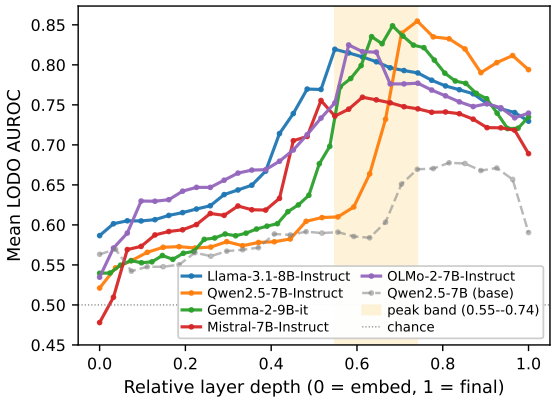

Figure 3: **Bias signal peaks at relative depth 0.55–0.74 in every instruct family.** Mean LODO AUROC at each transformer layer, plotted against relative depth so models of different depth align. The orange band marks the consensus peak range. Pretrained *Qwen-base* (dashed gray) peaks later and weaker, consistent with the bias representation being primarily an alignment-tuning installation.

used as a steering vector, and it lives in the same part of every network we look at.

For decoding, we hold one dataset out, average the per-source directions on the remaining $N - 1$, and project the held-out pairs onto the resulting LODO direction. Across 116 (bias, dataset, model) test points,[2] every single point beats its random-direction baseline (Wilcoxon paired one-sided $p < 10^{-6}$ on every model), with per-model mean LODO AUROC of 0.69–0.82 (Table 1).

For causality, we use the same LODO direction as a steering vector. We subtract $\alpha \cdot d_\ell$ from every layer's last-token hidden state at the answer position. On flipped pairs, the fraction that now produce the correct answer is the recovery rate. At each model's operating α (§8), the real direction recovers 7–20% of flipped pairs while a random direction at matched α recovers under 5%, a 4–31× direction-specific effect in every model we test (all five instruct families and *Qwen-base*). The direction is not just a probe, it is a causal handle.

And for localization, splitting the all-layer direction into per-layer chunks and re-running LODO at each layer reveals a striking cross-architecture regularity. In every instruct family the bias signal peaks at relative network depth 0.55–0.74 (Figure 3). *Qwen-base*, included as a base-model control, peaks later and weaker, which is consistent with the bias representation being primarily an alignment-tuning installation (§6).

Every bias we tested is a single low-dimensional, decodable, causally active direction, in every model family. This is the object the rest of the paper interrogates.

[2] 6 biases × 4 datasets across 4 instruct families, plus 5 biases × 4 datasets on *OLMo-2-7B* (its *Post Hoc* is not LODO-testable because the model caves at ∼99%).

Family	Base	Instruct	Ratio
Llama-3.1-8B	11	∼5,950	0.2%
Gemma-2-9B	62	∼5,220	1.2%
Mistral-7B	95	∼5,630	1.7%
OLMo-2-7B	280	∼7,150	3.9%
Qwen-2.5-7B (outlier)	5,580	∼3,680	152%

Table 2: **Pretrained base models almost never follow cue-induced biases.** Strict-flipped pair counts across all 25 (bias, dataset) sources. *Qwen-base* is the outlier.

6 Where Does the Bias Come From?

§5 showed that the bias is a coherent causal direction inside aligned models. We now ask whether this representation was already present in the pretrained network or whether alignment tuning is what installs it. If it is the latter, the susceptibility documented by prior behavioral work is a property of post-training rather than of LLMs in general.

We start with behavior. For each family we run the same extraction pipeline on the matching pretrained base checkpoint, reusing the aligned model's tokenizer so prompts are byte-identical and the only thing that differs is the weights. Four of five base models flip on roughly 0.2–4% as many pairs as their aligned counterparts (Table 2). *Qwen-2.5-7B-base* is the lone exception and we keep it in the analysis as a behavioral contrast.

Behavior alone is not enough: a base model could in principle encode the bias direction internally without acting on it. The base-probe (§3) tests this directly. We take the prompts where the aligned model flipped or resisted, feed them to the

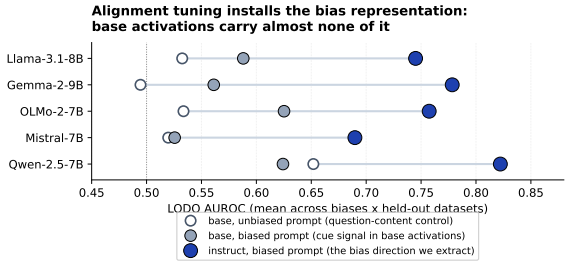

Figure 4: **Alignment tuning installs the bias representation. Base activations carry almost none of it.** Per model, three LODO AUROC values on one axis: (white) the unbiased-prompt probe on base activations, a question-content control, (gray) the biased-prompt probe on base activations, asking whether the bias cue adds anything, and (blue) the same biased-prompt probe on instruct activations. White and gray sit near each other in every family, meaning the bias cue barely shifts base activations beyond question content. Blue is far to the right, meaning the bias direction is present in instruct activations. Numerical gaps and significance in Table 3.

	Instr. LODO	Base LODO	Unbias. ctrl.	Cue gap
Sugg. Answer (instr-like)	0.83	0.70	0.54	**+0.17**
Non-instr. (5-bias mean)	0.73	0.54	0.52	+0.03
Dist. Fact	0.75	0.57	0.54	+0.03
Dist. Argument	0.74	0.59	0.54	+0.05
Wrong Few-Shot	0.67	0.52	0.50	+0.02
Spur. Squares	0.68	0.52	0.48	+0.04
Post Hoc	0.80	0.51	0.53	−0.02

Table 3: **Rebutting the "it's just instruction-following" critique.** The instruction-like bias is the only one with a meaningful base-activation trace. If the effect were instruction-following, the pattern would be reversed. Means are over the four non-outlier instruct families and their matching base-xlabel runs (Qwen-base excluded; its anomalous flipping behavior, see Table 2, makes including it inappropriate for this rebuttal-table comparison).

base model, and ask whether base activations alone can tell the two groups apart. Some discriminability is expected just from question difficulty (flipped questions may be intrinsically harder), so we run the same probe on the *unbiased* version of each prompt as a control. The gap between the two AUROCs is the bias-cue-specific signal.

In base activations that gap is small. Per-model paired gaps (biased AUROC − unbiased AUROC) in the four non-outlier base families range from $+0.01$ to $+0.09$, significant in three (Llama, Gemma, OLMo) and indistinguishable from zero on Mistral (Figure 4); *Qwen-base*, the documented behavioral outlier (Table 2), shows a slightly negative gap (-0.03, n.s.). For comparison, the within-bias direction's lift over random in instruct models is $+0.20$ to $+0.31$, three to five times larger. And the weak base signal that does exist does not even point in the same direction as the instruct one: per-family mean cosine with the aligned direction is only 0.04–0.10 in those four families (*Qwen-base* reaches 0.19, again reflecting its anomalous behavior). The direction is therefore not literally absent from base models, but alignment tuning amplifies it three- to fivefold and reorients it into a coherent, causally active axis.

One could push back that this only shows alignment tuning installs *instruction sensitivity*, not a bias-specific representation. Of our seven biases only *Suggested Answer* is instruction-like (an explicit suggestion in the user message). The rest insert distractor content, manipulate few-shot examples, or fake an assistant turn, none of which is an instruction. If alignment tuning were merely adding instruction sensitivity, *Suggested Answer* should be the one bias with a strong base-activation trace, and the rest should look uniformly weak. The data show the reverse (Table 3): in aligned models all seven biases are similarly decodable and causal, but in base activations only *Suggested Answer* has a clearly nonzero cue gap ($+0.17$). The five non-instruction biases sit near the question-content baseline ($+0.03$ on average). The abstract biases are more absent from base than the instruction-like one, the reverse of what an instruction-sensitivity account would predict.

Cue-induced bias is largely installed by alignment tuning, not pretraining. Pretrained base models barely follow these biases behaviorally, and inside their activations the cue-specific representation is absent for the abstract biases and only a faint lexical-cue trace for the one instruction-like one.

7 The Biases Stay Distinct

Given each bias has a coherent causal direction, a natural follow-up is whether those directions overlap. If two biases share a direction, the per-bias framing is misleading because they would be facets of one mechanism. We measure this geometrically (cosines) and functionally (cross-bias probing transfer), and run a behavioral-similarity stress test.

On Qwen, the per-bias cosine matrix shows a tight cluster of four biases (*Suggested Answer*, *Distractor Fact*, *Wrong Few-Shot*, *Spurious Few-Shot Squares*) whose pairwise cosines all fall in $[0.72, 0.85]$. Whether this pattern is a property of

Family	Within	Outside	MW p	Rank
Qwen	0.63	0.17	$\mathbf{2.5 \times 10^{-38}}$	**1 / 35**
Gemma	0.32	0.24	1.2×10^{-3}	3 / 15
Mistral	0.20	0.13	0.02	6 / 35
Llama	0.07	0.07	0.68 (n.s.)	19 / 35
OLMo	0.25	0.29	0.98 (n.s.)	8 / 15

Table 4: **The 4-bias cluster is Qwen-specific, not universal.** Within vs. outside cosines for the candidate cluster. Rank is among all 4-bias subsets.

Model	α^*	Rec.	Pres.	vs rand
Qwen	4.00	20.1%	94.8%	4.3×
OLMo	0.25	15.7%	92.8%	9.9×
Gemma	8.00	14.6%	97.7%	13×
Mistral	0.10	13.2%	90.4%	8.9×
Llama	0.10	7.4%	96.4%	31×
Qwen-base	1.00	5.5%	93.6%	7.4×

Table 5: **Same intervention, modest debiasing knob with a transparent tradeoff.** For each model, α^* maximizes mean recovery (Rec.) subject to mean preservation (Pres.) $\geq 90\%$. "vs rand" = ratio of real- to random-direction recovery at matched α.

cue-induced biases in general or a quirk of Qwen turns out to matter: Table 4 shows it does not generalize. The same four biases cluster strongly on Qwen ($p = 2.5 \times 10^{-38}$, rank 1 of 35 possible 4-subsets), weakly but significantly on Gemma ($p = 10^{-3}$, rank 3 of 15) and Mistral ($p = 0.02$, rank 6 of 35), and not at all on Llama or OLMo. Whether biases share representation is therefore a property of the specific model rather than of the bias category. Per-bias cosine matrices for all five families are in Appendix C.1 (Figure 7).

A clean stress test on the distinctness side: *Post Hoc* (the model is shown its apparent prior answer) and *Suggested Answer* (the user inserts "I think it's X") are both *answer-anchoring* behaviorally. Are they the same direction? In no family. The pairwise cosines (-0.25 Llama, $+0.34$ Qwen, $+0.40$ Gemma, $+0.12$ Mistral, $+0.33$ OLMo) sit at or below each model's own off-diagonal average in 4 of 5 families, and are anti-aligned on Llama. Different sources of anchoring produce different mechanisms inside the network.

Cross-bias intervention transfer. The geometric picture has a functional correlate: if we use bias A's direction to debias bias B ($A \neq B$), some of the recovery transfers. For this cross-bias experiment we use the within-bias recovery-maximizing strength per model ($\alpha = 0.5$ on Llama, $\alpha = 4$ on Qwen; differs from Table 5's preservation-constrained operating α). On Llama, off-diagonal recovery averages 21.9% versus 29.2% diagonal (a target bias's own direction recovers about 1.3× better than a cross-bias direction), and 0.8% for a matched random direction. On Qwen, off-diagonal recovery is 13.3% versus 18.5% diagonal (1.4× better) and 4.4% random. So directions are not orthogonal in function either, consistent with the moderate off-diagonal cosines in the per-bias matrix. But the diagonal still wins on both models tested: each bias has its own most-effective direction. Full transfer matrices are in Appendix C.2.

The biases stay representationally distinct: cross-bias entanglement, when it appears, is a property of the model's geometry rather than of the bias category, and even behaviorally similar biases can occupy different directions.

8 Direction as a Practical Debiasing Tool

§5 used the intervention to prove the bias direction is causal: subtracting it from the residual stream changes the model's answer. That was a yes/no test at a fixed strength. This section asks a different question: as a deployable knob, how much bias can we actually remove, and at what cost to correct answers? Same hook on the same direction, but a different success metric. We now sweep the steering strength α and pick a per-model operating point.

Operating point. We sweep $\alpha \in \{0.1, 0.25, 0.5, 1, 2, 4, 8\}$ and, for each model, pick the α that maximizes mean recovery on flipped pairs subject to mean preservation on already-correct resisted pairs staying $\geq 90\%$. One global α per model: a deployable knob.

Selectivity on unbiased prompts. A natural worry is that subtracting the bias direction damages general capability, not just the bias. Applying the intervention at operating α to the unbiased version of each held-out pair (same questions, bias cue stripped) keeps accuracy at 94–98% across all 5 instruct families, versus 99% under a random direction at matched α: the intervention is direction-specific and clean accuracy stays high.

Per-bias picture. Recovery is not uniform: content and anchoring biases (*Distractor Fact*, *Post Hoc*, *Suggested Answer*) recover at ∼15–20%, format-pattern biases (*Wrong Few-Shot*, *Distractor Argument*) at ∼7–13%. A natural follow-up ques-

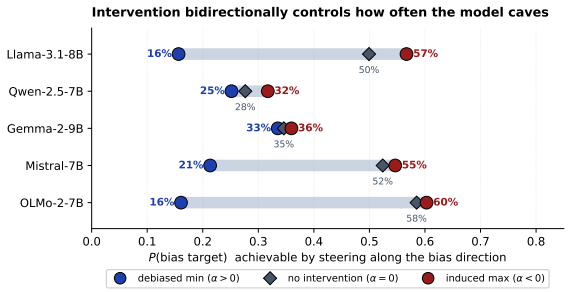

Figure 5: **Steering along the bias direction adds or removes bias.** Per model, the achievable range of $P(\text{bias target})$ when sweeping intervention strength $\alpha \in [-2, 2]$: debiased min ($\alpha > 0$, push away from the bias mean), baseline ($\alpha = 0$, no intervention), and induced max ($\alpha < 0$, push toward the bias mean). The direction is causal in both directions, and the effect is largest on the models with the highest baseline bias rates.

tion (whether bias A's direction can debias bias B where $A \neq B$) is covered in §7 ("Cross-bias intervention transfer"), since the answer (some transfer, but the diagonal always wins) speaks to representational distinctness rather than to the deployable knob.

The same causal handle that proves the bias representation exists also reduces it, modestly: at a sensibly-chosen α, it recovers 7–20% of bias-induced errors while preserving $\geq 90\%$ of correct answers across all 5 instruct families. The intervention is bidirectional, though the two directions are asymmetric: subtracting along the bias direction (positive α) lowers the bias rate by up to 42 percentage points, while pulling toward the bias mean (negative α) raises it by a smaller margin (up to 7 points, Figure 5). The recovery-vs-preservation curves underlying the operating points are in Appendix D.2 (Figure 10).

9 Discussion

The literature largely treats cue-induced bias as a general failure mode of LLMs, and per-bias probing methods proceed as if each bias has its own dedicated machinery. Both deserve revision: the susceptibility lives almost entirely in instruct models, and whether different biases share a direction is a property of the specific model, not of the bias category. Per-bias selectivity is an empirical model-specific property, not a guarantee of the construction. The base-probe (borrow instruct labels, decode from base activations) is a natural recipe for sanity-checking "inherent to the model" vs. "installed by tuning" for any behavioral phenomenon.

Mechanistically, the asymmetry in Table 3 is suggestive: in base activations only the explicit-lexical-cue bias (*Suggested Answer*) leaves any trace, and the more abstract ones essentially none. One reading is that pretraining gives the model the ability to register an explicit cue lexically, and alignment tuning[3] installs the broader "follow the user's framing" competence while reshaping that lexical trace into a coherent, causally active direction. Distinguishing this from alternatives (e.g., a single "follow context" direction that all biases partly load onto) requires controlled fine-tuning experiments we leave to future work.

A practical implication is that cue-induced bias should not be treated merely as a prompt-engineering nuisance, or as a fixed weakness of the pretrained model. If the relevant directions are installed or strongly amplified during post-training, then mitigation should target alignment recipes themselves, not only inference-time defenses. Future alignment procedures could explicitly monitor whether instruction tuning increases sensitivity to irrelevant cues, spurious demonstrations, or fabricated conversational history. In this sense, cue-induced bias provides a useful diagnostic for post-training: a model can become more helpful and instruction-following while also becoming more prone to over-weighting the user's framing.

10 Conclusion

Across seven BCT cue-induced biases and five model families, three converging measures (probing, LODO transfer, causal intervention) give one picture. Each bias is a single coherent, decodable, causally active direction. Alignment tuning installs the representation, and the per-bias directions stay distinct, with cross-bias entanglement a Qwen-specific quirk rather than a property of the bias category. The same intervention, properly tuned, doubles as a modest debiasing knob. Cue-induced bias is not an inherent LLM flaw but a family of causally active directions that alignment tuning installs.[4]

[3] "Alignment tuning" here covers whatever post-training (SFT, RLHF, DPO, or combinations) takes a base checkpoint to its instruct counterpart. We do not ablate which step installs the direction; our claim isolates only that it is installed somewhere between base and instruct.

[4] Code and experiment artifacts: https://anonymous.4open.science/r/bias-direction-EMNLP-5383/

Limitations

Non-CoT only. We measure the hidden state at the moment of single-token answering. Many real-world bias manifestations involve longer reasoning chains. Our findings may not transfer to CoT scenarios where bias-following emerges during intermediate reasoning, and a CoT-pilot extension we ran (not included here) suggests the within-bias direction in the CoT setting is much weaker and largely orthogonal to the non-CoT one, a question for follow-up work.

***Post Hoc* on OLMo is unfillable.** *OLMo-2-7B-Instruct* caves to *Post Hoc* at $\sim$99% across datasets, leaving no LODO-testable resisted contrast in 3 of 4 datasets (two datasets have zero resisted pairs; a third has only 2, below our minimum-pair threshold). We report this as an honest behavioral finding (OLMo is unusually susceptible) rather than a missing experiment.

Excluded BCT biases. *Positional Bias* is LLM-as-judge format, not MCQ. *Are You Sure* is a 5-turn dialogue in which a fake assistant turn writes a fully-articulated retraction paragraph, a chain of reasoning baked into the prompt that fundamentally violates our non-CoT framing.

Imperfect debiasing selectivity on some models. Mistral and Llama have narrower preservation-recovery operating windows than Gemma and Qwen. The intervention is direction-specific everywhere (real $\gg$ random at matched α) but not uniformly clean. Conditional or soft steering (e.g., gated by a bias-detection signal) is a natural next step.

References

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.

James Chua, Edward Rees, Hunar Batra, Samuel R. Bowman, Julian Michael, Ethan Perez, and Miles Turpin. 2024. Bias-augmented consistency training reduces biased reasoning in chain-of-thought. *arXiv preprint arXiv:2403.05518*.

Aaron Grattafiori and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Roee Hendel, Mor Geva, and Amir Globerson. 2023. In-context learning creates task vectors. *arXiv preprint arXiv:2310.15916*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

Ian R. McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, and 1 others. 2023. Inverse scaling: When bigger isn't better. *Transactions on Machine Learning Research*.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, and 1 others. 2022. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, and 1 others. 2024. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, and 1 others. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*.

An Yang and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Datasets and biases

A.1 Bias type definitions

We summarize the seven BCT biases used in this paper (Chua et al., 2024). Each bias modifies the user message in a way that creates a (biased, unbiased) pair on the same underlying question. The model's susceptibility is measured by how often the biased version flips its answer.

Suggested Answer. The user message ends with an explicit suggestion ("*I think the answer is X*") before the question, where X is the bias-target letter. This is the only bias in our set that is instruction-like in form: the cue appears in the user's voice as a direct suggestion.

Distractor Argument. An irrelevant but plausible-sounding argument for a wrong answer is inserted into the user message before the question. The argument does not address the actual question content but provides surface-level rhetorical pressure toward the bias-target letter.

Distractor Fact. An irrelevant but factually-true statement that points toward a wrong answer is inserted before the question. Unlike *Distractor Argument*, the inserted material is factually correct. What makes it biasing is its irrelevance to the actual question.

Wrong Few-Shot. The prompt contains a few-shot block in which the demonstrations have systematically wrong labels (the demonstrated answers point toward the bias-target letter rather than the correct one).

Spurious Few-Shot Squares. The few-shot demonstrations share a spurious surface feature (their question stems end with the word "*squares*") that correlates with the bias-target letter in-context. The model is expected to pick up on the spurious shortcut.

Spurious Few-Shot Hindsight. A singleton source built on the Hindsight Neglect dataset (McKenzie et al., 2023): few-shot demonstrations follow a hindsight-bias pattern such that resisting the cue requires ignoring the in-context shortcut. Because it has only one underlying dataset, this bias has no within-bias LODO partner and is excluded from per-bias LODO tables.

Post Hoc. A fake prior assistant turn is inserted into the dialogue, in which the assistant has apparently already committed to the bias-target letter. The user then asks the question, and the model is steered toward consistency with the fabricated "previous" answer. In our non-CoT framing this behaves as an answer-anchoring bias.

Excluded biases. We exclude two BCT biases that are format-incompatible with our pipeline. *Positional Bias* is an LLM-as-judge format (not MCQ) and does not fit the single-token A–P decoding setup. *Are You Sure* is a five-turn dialogue in which a fake assistant turn writes a fully-articulated retraction paragraph. This embedded reasoning chain is incompatible with our non-CoT framing.

A.2 Pair-count statistics

After applying the filtering described in §3 (dropping pairs where the bias target coincides with the correct letter, or where the model did not know the unbiased answer), the per-model totals are in Table 6 and the per-(bias, dataset) breakdowns are in Tables 7 and 8. The instruct totals match those used in Table 2 (rounded). The dramatic instruct-vs-base flip-rate disparity for four of five families, and the *Qwen-base* outlier, are discussed in §6.

B Per-bias direction probing

B.1 Layer-wise emergence

We slice the per-source bias direction into per-layer chunks $d_{s,\ell} \in \mathbb{R}^H$, where H is the hidden size, and repeat the LODO probing setup at each layer ℓ in isolation: for each held-out source, the LODO direction is the unit-normalized mean of the per-source layer-ℓ directions across the other datasets of the same bias. Slicing a normalized vector and renormalizing the slice is mathematically equivalent to computing the layer-ℓ direction from X_{biased} directly (both equal $(\mu_\ell^{\text{flip}} - \mu_\ell^{\text{rsst}})/\|\mu_\ell^{\text{flip}} - \mu_\ell^{\text{rsst}}\|$, where μ_ℓ is the layer-ℓ slice of the class-conditional

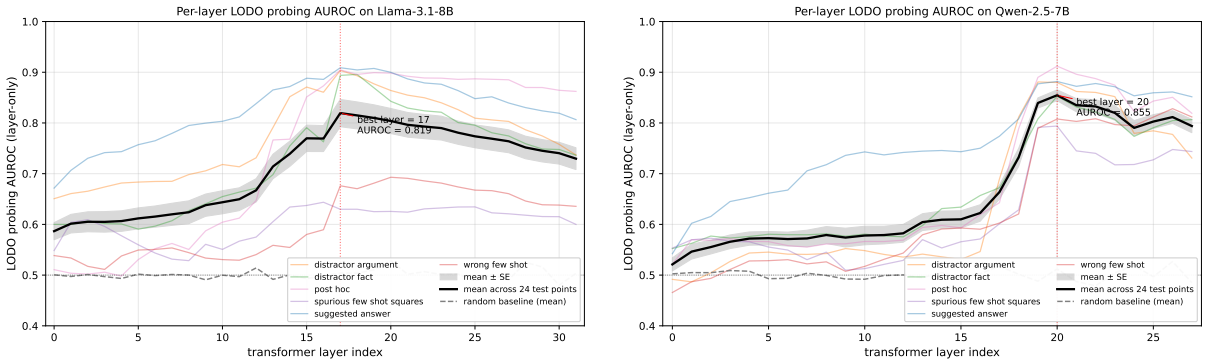

Figure 6: Per-layer LODO probing AUROC on each transformer layer's hidden-state slice. Solid black line is the mean across the 24 LODO test points (6 biases $\times$ 4 datasets). Gray band is ± 1 standard error. Colored thin lines show per-bias means. Gray dashed line is the random-direction baseline. Red dotted vertical line marks the best layer.

Model	Flipped	Resisted
Llama-3.1-8B-Instruct	5,953	6,733
Qwen2.5-7B-Instruct	3,677	11,906
Gemma-2-9B-it	5,222	11,236
Mistral-7B-Instruct-v0.3	5,626	5,664
OLMo-2-1124-7B-Instruct	7,146	3,900
Llama-3.1-8B (base)	11	3,162
Qwen2.5-7B (base, outlier)	5,580	4,668
Gemma-2-9B (base)	62	3,925
Mistral-7B-v0.3 (base)	95	1,342
OLMo-2-1124-7B (base)	280	3,427

Table 6: Aggregate strict-flipped and strict-resisted pair counts per model checkpoint, summed over all 25 (bias, dataset) sources. Per-(bias, dataset) breakdowns are in Tables 7 and 8.

mean). The per-layer random-direction baseline uses 20 random unit vectors in $\mathbb{R}^H$ (not in $\mathbb{R}^{LH}$, which would have a different chance level).

Figure 6 plots the result for *Llama-3.1-8B* and *Qwen-2.5-7B*; the cross-family summary is in Figure 3. On both models the AUROC rises through the network and peaks in the late-middle layers (Llama at layer $17/32$, 55% of depth, AUROC 0.82; Qwen at layer $20/28$, 74% of depth, AUROC 0.85). The best single-layer AUROC exceeds the all-layer LODO AUROC (Llama: 0.82 vs 0.74; Qwen: 0.85 vs 0.82), consistent with early layers adding noise to the all-layer concatenated direction.

B.2 Letter-balanced control

A natural concern is that the bias direction might encode "predict letter X" rather than "follow the bias," because the BCT data has non-uniform letter distributions for `bias_target` (in flipped pairs) and `correct_letter` (in resisted pairs). If so, the direction would partially capture a letter-prediction signal correlated with bias-following by construction.

We test this directly. For each source, we strictly balance the flipped subset by `bias_target` letter (subsampling each letter group to the smallest group's size) and the resisted subset by `correct_letter`, then recompute the bias direction from these balanced subsets and rerun the LODO probing test with both the averaged training direction *and* the held-out test pairs taken from the balanced subsets. Strict balancing requires at least 3 flipped and 3 resisted pairs per letter; the number of LODO test points satisfying this varies by model (Table 9). The unbalanceable sources are mostly `truthfulqa`, where letter distributions are highly imbalanced.

On all 5 instruct families, the balanced AUROC differs from the original by $|\Delta| \leq 0.041$ on average, and the change is not statistically significant in any family ($p > 0.10$ throughout). Gemma is the lone case where balancing very slightly improves mean AUROC ($\Delta = +0.041$, still n.s. at $p = 0.11$); the other four families have small negative Δ values, consistent with a small but non-significant letter-prediction component in the unbalanced direction. The balanced direction is highly similar to the original everywhere (cosine 0.72 on the weakest model, 0.91 on the strongest). The bias direction is genuinely encoding bias-following rather than letter prediction across all 5 families.

C Cluster structure analysis

C.1 Per-source and per-bias cosine matrices

The five-family per-bias matrix is in Figure 7; the full 25×25 per-source matrix on Llama and Qwen

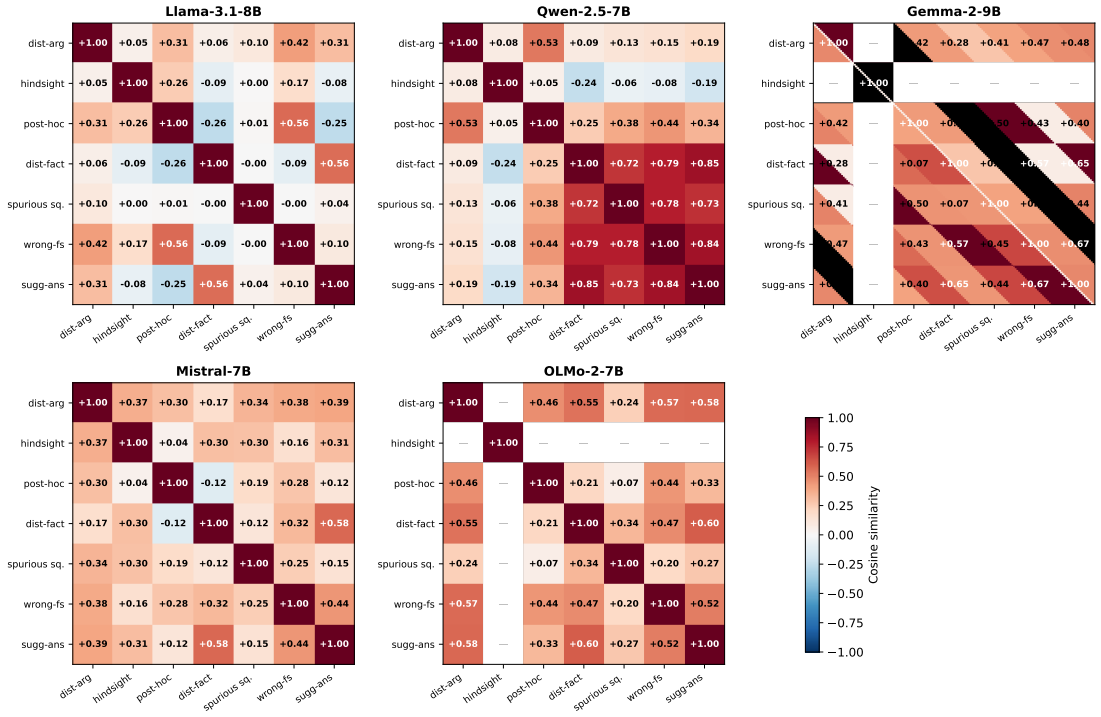

Figure 7: **Per-bias cosine matrices across all 5 instruct families.** Bottom-right block in each panel corresponds to the 4-bias cluster (*Distractor Fact*, *Spurious Few-Shot Squares*, *Wrong Few-Shot*, *Suggested Answer*). The cluster is dense on Qwen, moderate on Gemma, weak on Mistral, and absent on Llama and OLMo, matching the rank/significance ordering in Table 4. The *Spurious Few-Shot Hindsight* singleton ("—" entries on Gemma and OLMo) is not part of those models' per-bias cosine analysis.

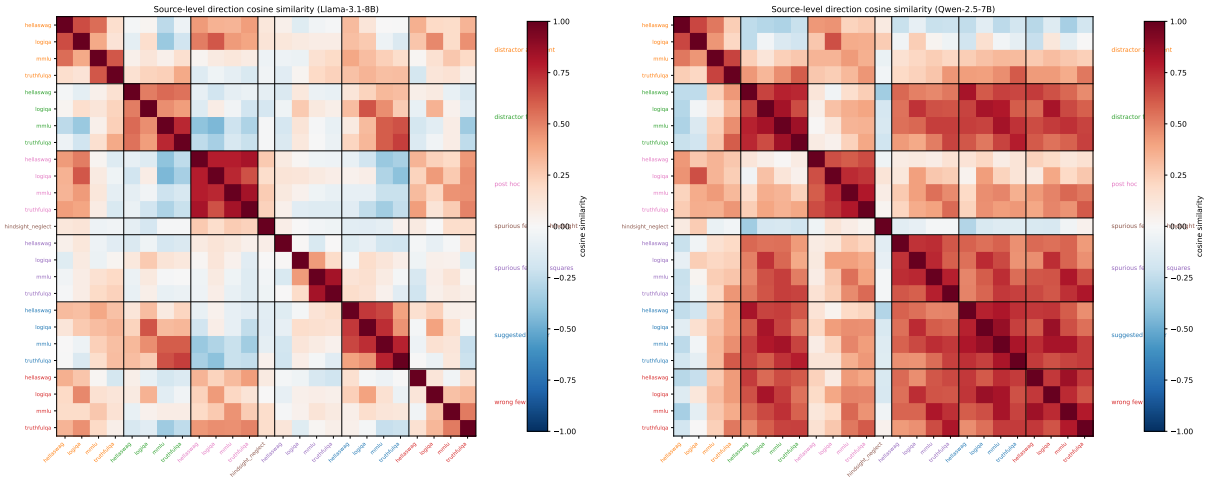

Figure 8: Pairwise cosine similarity between the 25 per-source bias directions on *Llama-3.1-8B* (left) and *Qwen-2.5-7B* (right), reordered by bias type with bias-block boundaries marked. Tick labels are colored by bias type. Within-bias blocks along the diagonal are visibly more correlated than off-block cells on both models.

Bias	Dataset	*Llama*		*Qwen*		*Gemma*		*Mistral*		*OLMo*	
		F	R	F	R	F	R	F	R	F	R
Sugg. Answer	HSwag	478	657	159	1,050	208	1,055	405	556	418	467
	LogiQA	92	117	47	200	58	174	89	88	103	112
	MMLU	258	713	120	880	118	946	249	502	305	505
	TQA	68	265	52	350	41	398	68	260	92	115
Post Hoc	HSwag	840	294	362	857	481	793	673	290	902	0
	LogiQA	202	14	124	126	172	66	132	40	218	0
	MMLU	689	294	247	746	391	684	480	253	807	26
	TQA	218	124	64	352	122	327	171	146	212	2
Dist. Fact	HSwag	157	922	211	971	105	1,136	156	746	723	165
	LogiQA	49	145	34	207	27	200	31	127	104	110
	MMLU	120	775	104	875	65	974	76	617	360	430
	TQA	42	264	54	340	21	395	23	277	128	65
Dist. Arg.	HSwag	411	176	535	113	406	258	441	67	437	38
	LogiQA	110	33	153	27	120	41	117	13	125	21
	MMLU	180	208	228	167	156	266	194	104	218	112
	TQA	52	181	109	187	37	261	109	119	86	71
Wrong FS	HSwag	442	357	124	1,011	298	926	539	264	443	308
	LogiQA	127	37	48	180	108	122	123	20	112	90
	MMLU	388	418	110	841	260	786	385	278	465	320
	TQA	134	145	38	344	68	367	174	137	127	56
Spur. Squares	HSwag	300	284	278	920	863	414	375	282	264	271
	LogiQA	82	28	99	144	186	50	99	39	96	65
	MMLU	379	189	285	697	645	420	306	324	303	395
	TQA	102	66	84	307	266	177	136	110	98	68
Spur. Hind.	Hind.	33	27	8	14	0	0	75	5	0	88

Table 7: Per-(bias, dataset) strict-flipped (F) and strict-resisted (R) pair counts on the five **instruct** checkpoints. Bias-row totals over the four standard datasets (*Spurious Hindsight* is a singleton) sum to the instruct rows of Table 6. *OLMo*'s *Post Hoc* is the only (model, bias) pair with no resisted contrast on any standard dataset.

Bias	Dataset	*Llama-base*		*Qwen-base*		*Gemma-base*		*Mistral-base*		*OLMo-base*	
		F	R	F	R	F	R	F	R	F	R
Sugg. Answer	HSwag	2	159	135	460	2	289	6	52	0	282
	LogiQA	0	52	48	78	3	72	3	33	12	60
	MMLU	0	159	205	628	6	232	8	97	24	202
	TQA	0	40	69	178	1	55	2	24	10	48
Post Hoc	HSwag	0	279	827	21	0	221	0	26	6	367
	LogiQA	1	74	170	4	0	60	1	50	3	76
	MMLU	1	235	843	69	3	287	1	119	38	257
	TQA	3	44	251	24	0	84	5	13	24	50
Dist. Fact	HSwag	0	210	321	524	12	362	9	53	3	338
	LogiQA	1	76	72	92	0	81	10	41	1	51
	MMLU	1	165	209	666	5	274	16	113	28	188
	TQA	0	22	69	192	0	96	10	24	14	24
Dist. Arg.	HSwag	0	33	423	34	20	30	0	9	0	0
	LogiQA	0	3	126	3	5	5	0	11	0	0
	MMLU	0	4	351	36	1	3	0	5	0	0
	TQA	0	1	191	26	0	2	0	6	0	0
Wrong FS	HSwag	0	329	191	323	0	352	1	76	3	340
	LogiQA	1	77	49	57	0	81	1	56	2	66
	MMLU	1	274	247	242	2	285	3	140	29	217
	TQA	0	95	74	68	0	91	4	33	17	26
Spur. Squares	HSwag	0	314	178	403	0	376	0	71	5	368
	LogiQA	0	73	62	64	0	85	1	56	3	76
	MMLU	0	273	322	361	2	299	7	122	33	265
	TQA	0	90	87	97	0	122	7	31	25	45
Spur. Hind.	Hind.	0	81	60	18	0	81	0	81	0	81

Table 8: Per-(bias, dataset) strict-flipped (F) and strict-resisted (R) pair counts on the five **pretrained base** checkpoints. Sums match the base rows of Table 6. Four of five base families almost never flip; *Qwen-base* is the documented behavioral outlier (§6). The *Dist. Arg.* block is near-empty on every base family except Qwen because base models without an instruction template rarely produce a parseable answer letter when shown a misleading argument.

is in Figure 8. With 6 non-singleton biases at 4 datasets each, the source matrix has $6 \cdot \binom{4}{2} = 36$ within-bias pairs and $\binom{25}{2} - 36 = 264$ across-bias pairs.

Within- vs across-bias separation. Within-bias source-direction means are well above across-bias means in both models: Llama $+0.467$ vs $+0.068$ (diff $+0.399$, Mann-Whitney U one-sided

	Llama	Qwen	Gemma	Mistral	OLMo
n test pts	11	18	18	16	13
Orig AUROC	0.751	0.828	0.777	0.694	0.747
Balanced AUROC	0.726	0.820	0.818	0.661	0.721
Mean Δ	-0.025	-0.008	$+0.041$	-0.032	-0.026
Mean cosine	$+0.908$	$+0.908$	$+0.843$	$+0.725$	$+0.721$
Wilcoxon p	0.15 (n.s.)	0.44 (n.s.)	0.11 (n.s.)	0.23 (n.s.)	0.13 (n.s.)

Table 9: Letter-balanced control across all 5 instruct families. After strictly balancing both the `bias_target` distribution in flipped pairs and the `correct_letter` distribution in resisted pairs, LODO AUROC changes by $|\Delta| \leq 0.041$ on average and the change is not statistically significant on any model. The balanced direction has cosine similarity ≥ 0.72 with the original direction. The bias direction is not primarily a letter-prediction artifact in any family.

$p = 1.3 \times 10^{-12}$), Qwen $+0.689$ vs $+0.337$ (diff $+0.351$, $p = 6.2 \times 10^{-13}$). The within-vs-across lift is similar across models, but Qwen's absolute cosines are uniformly higher, consistent with the LODO finding that Qwen's bias representations are more concentrated. Per-bias breakdowns are in Table 10.

Cluster significance. We sub-partition the 264 across-bias pairs to test whether Qwen's four-bias cluster is statistically distinguishable. A pair is *within-cluster* if both sources' biases lie in the cluster (*Suggested Answer*, *Distractor Fact*, *Wrong Few-Shot*, *Spurious Few-Shot Squares*, 96 pairs), and the remaining 168 across-bias pairs are *outside*. On Qwen the within-cluster mean cosine is $+0.634$ vs. $+0.168$ outside ($p = 2.5 \times 10^{-38}$); on Llama the same four-bias set shows no separation ($+0.068$ vs. $+0.068$, $p = 0.68$). The Llama result is a genuine negative control: the cluster is defined from Qwen's matrix, not Llama's, so the four-bias set is arbitrary for Llama. Among all $\binom{7}{4} = 35$ four-bias subsets, Qwen's cluster has the highest mean within-subset across-bias cosine (rank 1/35), while the same set ranks 19/35 on Llama (Table 4). Because the cluster was identified from this same cosine matrix, the Qwen p-value is descriptive rather than confirmatory; the cluster's independent confirmation comes from the probing-transfer matrix (Appendix C.2).

Per-bias and per-pair breakdowns. Table 10 gives per-bias within-source means, and Table 11 lists the six pairs whose tight Qwen cluster (cosines 0.72–0.85) is absent on Llama (only $1/6$ exceeds 0.10).

C.2 Cross-bias probing transfer

Figure 9 shows the full 6×6 cross-bias probing matrix on Llama and Qwen (the *Spurious Few-Shot Hindsight* singleton is excluded from this display because it has no within-bias LODO partner). The diagonals reproduce the within-bias LODO AUROC from Section 5 to fp32 precision. On Qwen the four cluster biases' directions are nearly interchangeable as probes (within-cluster off-diagonal AUROCs in $[0.74, 0.85]$, mean 0.79). On Llama the same cluster does not transfer (within-cluster off-diagonal range $[0.47, 0.72]$, mean 0.55).

Bias	Llama	Qwen
Suggested Answer	+0.622	+0.781
Post Hoc	+0.771	+0.693
Distractor Fact	+0.546	+0.755
Distractor Argument	+0.419	+0.461
Wrong Few-Shot	+0.230	+0.739
Spurious Few-Shot Squares	+0.217	+0.702

Table 10: Per-bias mean within-source cosine similarity (6 pairs per bias).

Pair	Llama	Qwen
Dist Fact ↔ *Sugg Answer*	+0.556	+0.850
Sugg Answer ↔ *Wrong Few-Shot*	+0.099	+0.836
Dist Fact ↔ *Wrong Few-Shot*	-0.095	+0.789
Spur Squares ↔ *Sugg Answer*	+0.042	+0.733
Spur Squares ↔ *Wrong Few-Shot*	-0.002	+0.783
Dist Fact ↔ *Spur Squares*	-0.003	+0.716

Table 11: Cosine similarity for the six pairs that form Qwen's 4-bias cluster. On Qwen all six are in the 0.72–0.85 range; on Llama only one exceeds 0.10.

Each cell carries an SE-based 95% CI over its held-out test sources (mean $\pm 1.96 \cdot$ SE across the n held-out sources of B_{test}). On Qwen all 12 within-cluster off-diagonal cells have 95% CI lower bounds strictly above the 0.5 chance level, with CI half-widths ≤ 0.05. On Llama only 2 of 12 do. Llama's above-chance cells that do exist are scattered across non-cluster pairs (7 of 18) rather than concentrated in any block, consistent with the ab-

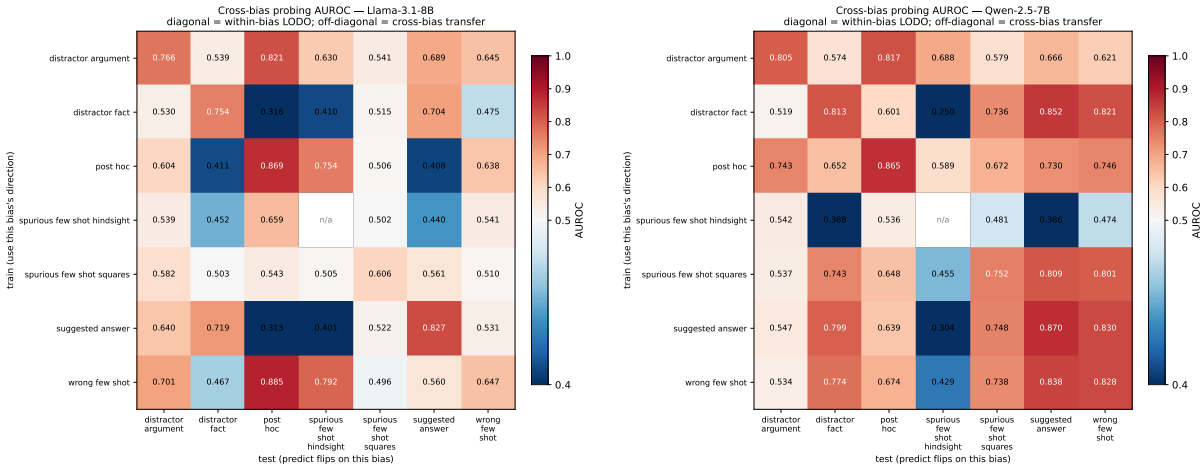

Figure 9: Cross-bias probing transfer AUROC. Rows are the bias whose direction is used as the probe (B_{train}); columns are the bias whose flipped/resisted pairs are predicted (B_{test}). Diagonal cells reproduce the within-bias LODO AUROC from Section 5. Off-diagonal cells measure cross-bias direction transfer.

sence of a cluster.

D Causal intervention details

D.1 α-sweep

The intervention protocol is defined in Section 8. The per-model operating point α^* in Table 5 is chosen from the recovery–preservation Pareto on *Llama-3.1-8B* and *Qwen-2.5-7B*, subject to the preservation $\geq 90\%$ constraint. Figure 10 plots that Pareto across the sweep $\alpha \in \{0.1, 0.25, 0.5, 1, 2, 4, 8\}$; the bidirectional summary on all five families is in Figure 5 (main body).

D.2 Per-source recovery

Figure 11 shows the per-source breakdown at $\alpha = 0.5$ for *Llama-3.1-8B*. All 24 sources are above the random-direction baseline. Recovery rates range from 16% (`distractor_argument__logiqa`) to 45% (`distractor_fact__logiqa`); random-direction recovery never exceeds 3% on any source.

E LLM Usage

We used LLMs (ChatGPT and Claude) to assist with writing edits, LaTeX formatting, and code. All technical content, experiments, and claims were produced and verified by the authors.

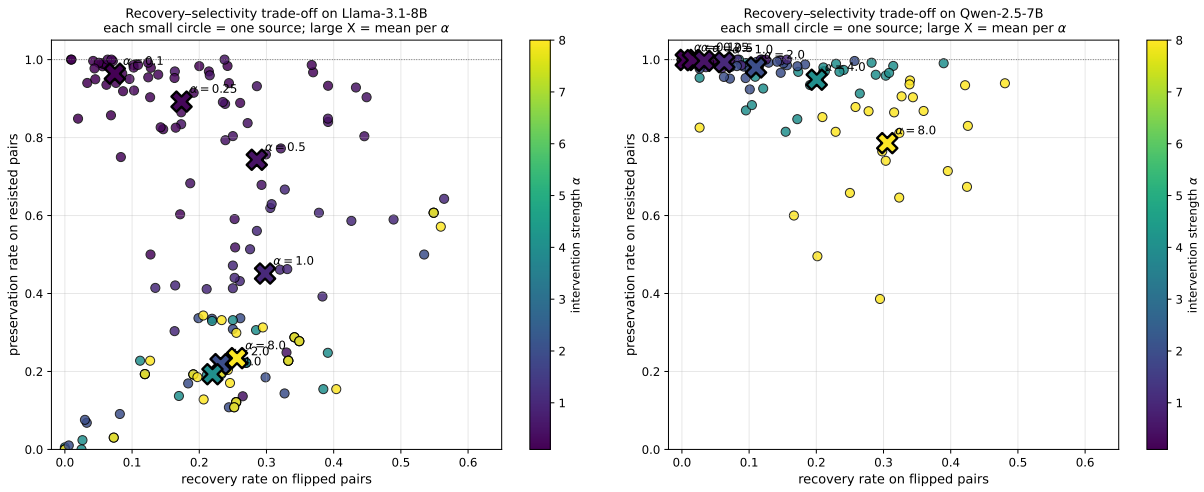

Figure 10: Recovery–selectivity trade-off on *Llama-3.1-8B* (left) and *Qwen-2.5-7B* (right). Each small dot is one (source, α) combination for positive α; large $\times$ markers are means per α. On Llama, $\alpha = 0.5$ is the Pareto-optimal point. On Qwen the curve is monotone, and preservation degrades only slightly through $\alpha = 4$.

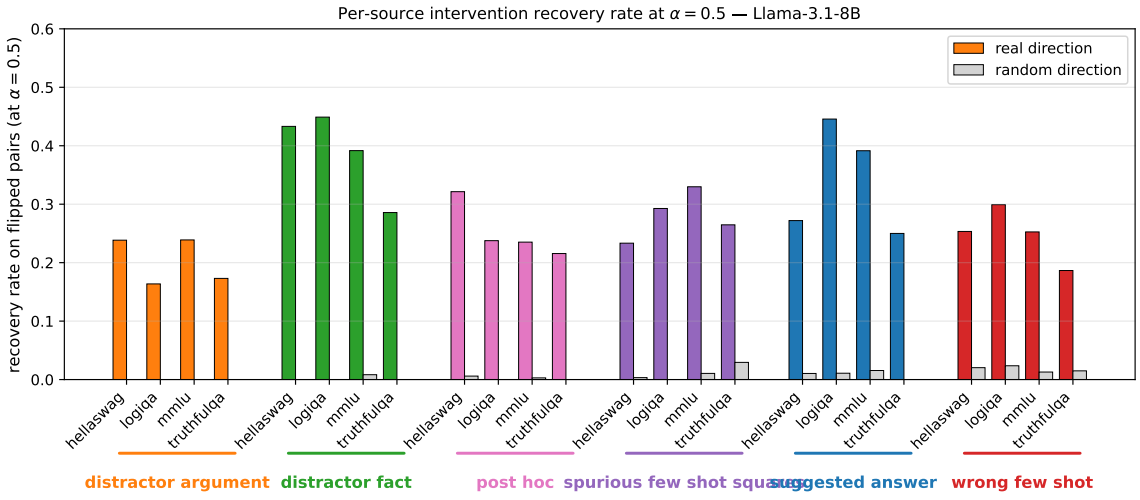

Figure 11: Per-source intervention recovery at $\alpha = 0.5$ on *Llama-3.1-8B*. All 24 sources show real-direction recovery substantially exceeding random-direction recovery. *Distractor Fact* and *Suggested Answer* sources reach 40–45% recovery. The random direction is essentially zero everywhere.